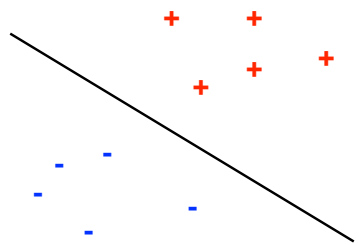
# 10-701 Introduction to Machine Learning (PhD)
# Lecture 11: SVMs

Leila Wehbe
Carnegie Mellon University
Machine Learning Department

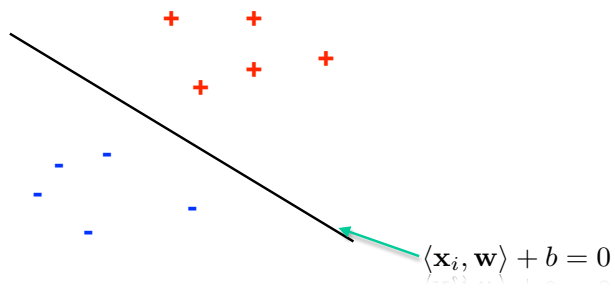
Readings: Andrew Ng's lecture notes at:
http://cs229.stanford.edu/notes/cs229-notes3.pdf

## SVMs

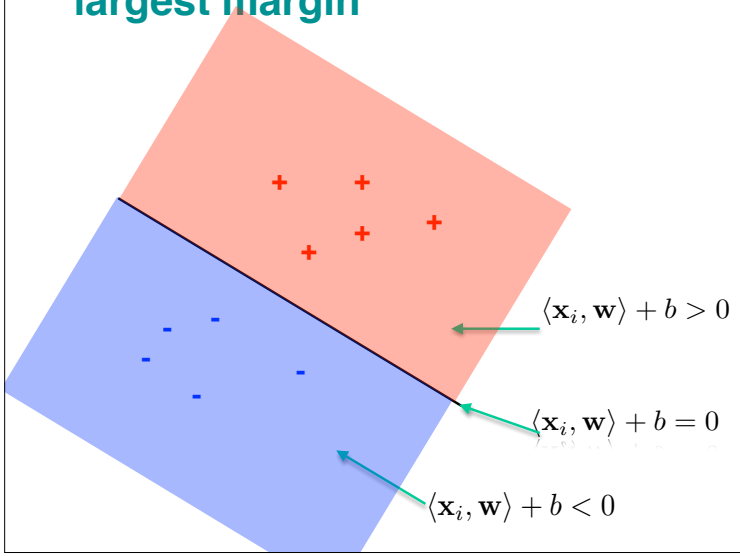
## Find a linear separator with the largest margin



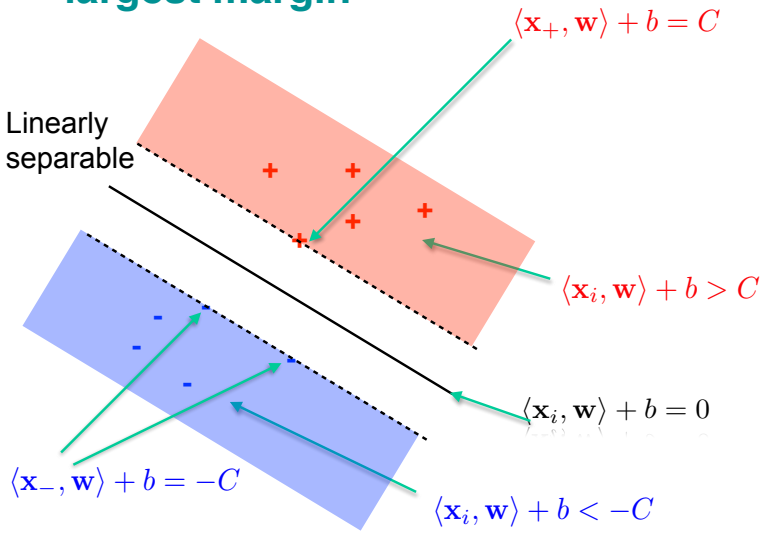
## Find a linear separator with the largest margin



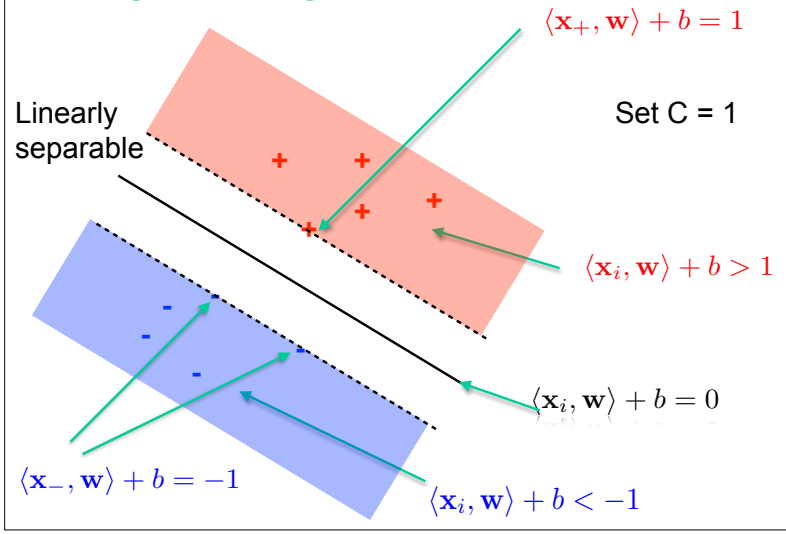
## Find a linear separator with the largest margin



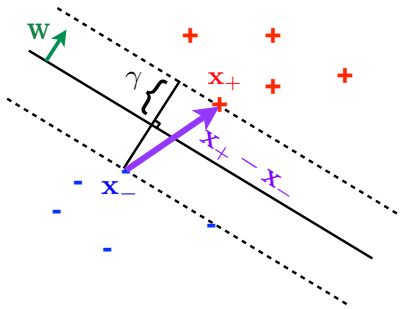

## Find a linear separator with the largest margin




## Find a linear separator with the largest margin




## Find a linear separator with the largest margin

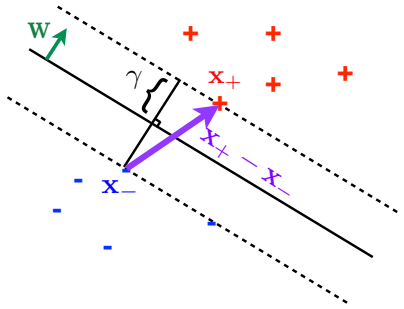



$$2\gamma = \frac{\langle \mathbf{x}_+ - \mathbf{x}_-, \mathbf{w} \rangle}{||\mathbf{w}||}$$

$$\gamma = \frac{1}{2} \frac{\langle \mathbf{x}_+ - \mathbf{x}_-, \mathbf{w} \rangle}{||\mathbf{w}||}$$

## Find a linear separator with the largest margin

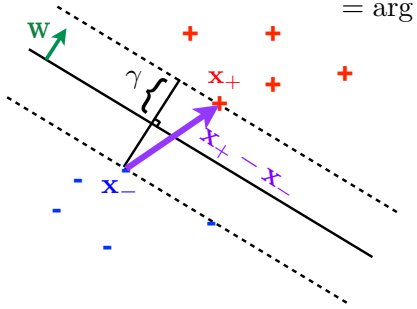
$$\arg\max_{\mathbf{w},b} \gamma = \arg\max_{\mathbf{w},b} \frac{1}{2} \frac{\langle \mathbf{x}_+ - \mathbf{x}_-, \mathbf{w} \rangle}{||\mathbf{w}||}$$




## Find a linear separator with the largest margin

$$\arg\max_{\mathbf{w},b} \gamma = \arg\max_{\mathbf{w},b} \frac{1}{2} \frac{\langle \mathbf{x}_+ - \mathbf{x}_-, \mathbf{w} \rangle}{||\mathbf{w}||}$$

$$= \arg\max_{\mathbf{w},b} \frac{1}{2} \frac{\langle \mathbf{x}_+, \mathbf{w} \rangle - \langle \mathbf{x}_-, \mathbf{w} \rangle}{||\mathbf{w}||}$$




$\langle \mathbf{x}_+, \mathbf{w} \rangle + b = 1$

$\langle \mathbf{x}_-, \mathbf{w} \rangle + b = -1$

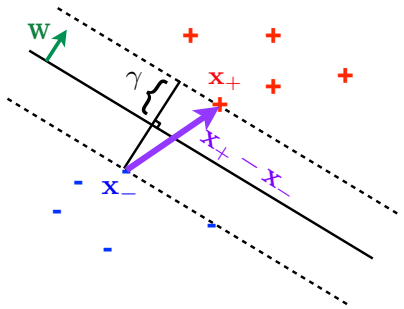
## Find a linear separator with the largest margin

$$\arg\max_{\mathbf{w},b} \gamma = \arg\max_{\mathbf{w},b} \frac{1}{2} \frac{1 - b - (-1 - b)}{||\mathbf{w}||}$$

$$= \arg\max_{\mathbf{w},b} \frac{1}{2} \frac{2}{||\mathbf{w}||} = \arg\max_{\mathbf{w},b} \frac{1}{||\mathbf{w}||}$$

$$= \arg\min_{\mathbf{w},b} ||\mathbf{w}||$$

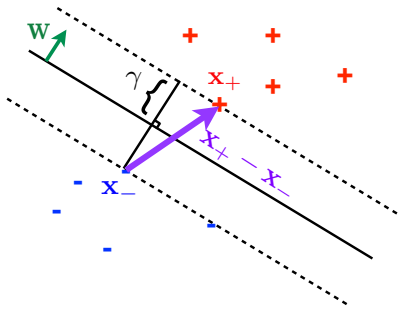
$$= \arg\min_{\mathbf{w},b} \frac{1}{2} ||\mathbf{w}||^2$$




## The (primal) optimization problem is:

$$\min_{\mathbf{w},b} \frac{1}{2} ||\mathbf{w}||^2$$

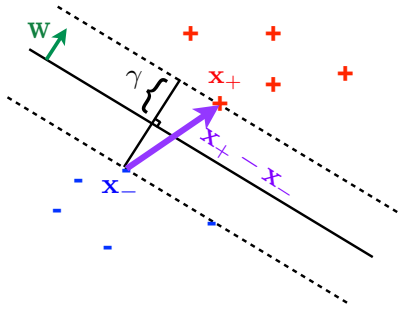
$$\text{s.t.} \quad y_i(\langle \mathbf{x}_i, \mathbf{w} \rangle + b) \geq 1, \quad i = 1, ..., m$$



## The (primal) optimization problem is:

$$\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2$$

$$\text{s.t.} \quad y_i(\langle \mathbf{x}_i, \mathbf{w} \rangle + b) \geq 1, \quad i = 1, ..., m$$




This can be written as:

$$\min_{\mathbf{w},b} \; f(\mathbf{w}, b)$$

$$\text{s.t.} \quad g(\mathbf{w}, b) \leq 0, \quad i = 1, ..., m$$

Where $f(\mathbf{w}, b) = ||\mathbf{w}||^2$ and $g(\mathbf{w}, b) = 1 - y_i(\langle \mathbf{x}_i, \mathbf{w} \rangle + b)$ are both convex functions of our parameters **w** and b

## Lagrangian

**We can write the Lagrangian of:**

$$\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2$$

$$\text{s.t.} \quad y_i(\langle \mathbf{x}_i, \mathbf{w} \rangle + b) \geq 1, \quad i = 1, ..., m$$

**as:** $\mathcal{L}(\mathbf{w}, b, \alpha) = f(\mathbf{w}, b) + \sum_{i=1}^{m} \alpha_i g_i(\mathbf{w}, b)$

$$= \frac{1}{2}||\mathbf{w}||^2 + \sum_i \alpha_i \left(1 - y_i(\langle \mathbf{x_i}, \mathbf{w} \rangle + b)\right)$$

$$g_i(\mathbf{w}, b) = 1 - y_i(\langle \mathbf{x}_i, \mathbf{w} \rangle + b)$$

## Lagrangian

**We can write the Lagrangian of:**

$$\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2$$

$$\text{s.t.} \quad y_i(\langle \mathbf{x}_i, \mathbf{w} \rangle + b) \geq 1, \quad i = 1, ..., m$$

**as:** $\mathcal{L}(\mathbf{w}, b, \alpha) = f(\mathbf{w}, b) + \sum_{i=1}^{m} \alpha_i g_i(\mathbf{w}, b)$

$$= \frac{1}{2}||\mathbf{w}||^2 + \sum_i \alpha_i \left(1 - y_i(\langle \mathbf{x_i}, \mathbf{w} \rangle + b)\right)$$

**Take the quantity:**

$$\theta_{\mathcal{P}}(\mathbf{w}, b) = \max_{\alpha, \alpha_i \geq 0} \mathcal{L}(\mathbf{w}, b, \alpha)$$

$$\theta_{\mathcal{P}}(\mathbf{w}, b) = \begin{cases} f(\mathbf{w}, b) & \text{if } \mathbf{w}, b \text{ satisfy primal constraint} \\ \infty & \text{otherwise} \end{cases}$$

## Primal and dual

$$\theta_{\mathcal{P}}(\mathbf{w}, b) = \max_{\alpha, \alpha_i \geq 0} \mathcal{L}(\mathbf{w}, b, \alpha)$$

$$\theta_{\mathcal{P}}(\mathbf{w}, b) = \begin{cases} f(\mathbf{w}, b) & \text{if } \mathbf{w}, b \text{ satisfy primal constraint} \\ \infty & \text{otherwise} \end{cases}$$

**The solution to:**

$$\min_{\mathbf{w},b} \theta_{\mathcal{P}}(\mathbf{w}, b) = \min_{\mathbf{w},b} \max_{\alpha, \alpha_i \geq 0} \mathcal{L}(\mathbf{w}, b, \alpha)$$

**This is the same problem as our original primal problem and therefore has the same solution.**
**Original problem:**

$$\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2$$

$$\text{s.t.} \quad y_i(\langle \mathbf{x}_i, \mathbf{w} \rangle + b) \geq 1, \quad i = 1, ..., m$$

## Primal and dual

**Now take a slightly different problem:**

$$\theta_{\mathcal{D}}(\alpha) = \min_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b, \alpha)$$

**The dual optimization problem is:**

$$\max_{\alpha,\alpha_i \geq 0} \theta_{\mathcal{D}}(\alpha) = \max_{\alpha,\alpha_i \geq 0} \min_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b, \alpha)$$

**We have that:**

$$d^* = \max_{\alpha,\alpha_i \geq 0} \min_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b, \alpha) \leq \min_{\mathbf{w},b} \max_{\alpha,\alpha_i \geq 0} \mathcal{L}(\mathbf{w}, b, \alpha) = p^*$$

**Under certain conditions (next slide):** $d^* = p^*$

## Primal and dual

Both satisfied in separable SVM formulation

If:
- f and g are convex
- There exist w that satisfies constraint g (i.e. there exists a w for which g(w)<0)

then:
- there exists **w*** and b* solutions to the primal problem, and $\alpha^*$ solution to the dual problem, and:

$$d^* = p^* = \mathcal{L}(\mathbf{w}^*, b^*, \alpha^*)$$

Also, $\mathbf{w}^*, b^*, \alpha^*$ satisfy the KKT conditions:

$$\left.\frac{\partial \mathcal{L}(\mathbf{w}, b^*, \alpha^*)}{\partial w_i}\right|_{\mathbf{w}^*} = 0, \ i = 1, ..., n$$

$$\left.\frac{\partial \mathcal{L}(\mathbf{w}^*, b, \alpha^*)}{\partial b}\right|_{b^*} = 0$$

$$\alpha_i^* g_i(\mathbf{w}^*) = 0, \ i = 1, ..., m$$

$$g_i(\mathbf{w}^*) \leq 0, \ i = 1, ..., m$$

$$\alpha_i \geq 0, \ i = 1, ..., m$$

## Primal and dual

Both satisfied in separable SVM formulation

If:
- f and g are convex
- There exist w that satisfies constraint g (i.e. there exists a w for which g(w)<0)

then:
- there exists **w*** and b* solutions to the primal problem, and $\alpha^*$ solution to the dual problem, and:

$$d^* = p^* = \mathcal{L}(\mathbf{w}^*, b^*, \alpha^*)$$

Also, $\mathbf{w}^*, b^*, \alpha^*$ satisfy the KKT conditions:

$$\left.\frac{\partial \mathcal{L}(\mathbf{w}, b^*, \alpha^*)}{\partial w_i}\right|_{\mathbf{w}^*} = 0, \ i = 1, ..., n$$

$$\left.\frac{\partial \mathcal{L}(\mathbf{w}^*, b, \alpha^*)}{\partial b}\right|_{b^*} = 0$$ Stationarity

Complementary slackness

$$\alpha_i^* g_i(\mathbf{w}^*) = 0, \ i = 1, ..., m$$

$$g_i(\mathbf{w}^*) \leq 0, \ i = 1, ..., m$$

Feasibility $\alpha_i \geq 0, \ i = 1, ..., m$

## KKT conditions

Stationarity: **w***,b* local extremum of Lagrangian for fixed $\alpha^*$

Feasibility: All primal and dual constraints are satisfied

Complementary Slackness: either $\alpha_i = 0$ or $g_i(\mathbf{w}, b) = 0$

If $\alpha_i > 0$ then $g_i(\mathbf{w}, b) = 0$

$$\left.\frac{\partial \mathcal{L}(\mathbf{w}, b^*, \alpha^*)}{\partial w_i}\right|_{\mathbf{w}^*} = 0, \ i = 1, ..., n$$

$$\left.\frac{\partial \mathcal{L}(\mathbf{w}^*, b, \alpha^*)}{\partial b}\right|_{b^*} = 0$$ Stationarity

Complementary slackness

$$\alpha_i^* g_i(\mathbf{w}^*) = 0, \ i = 1, ..., m$$

$$g_i(\mathbf{w}^*) \leq 0, \ i = 1, ..., m$$

Feasibility $\alpha_i \geq 0, \ i = 1, ..., m$

## Solve the dual problem. First solve:

$$\min_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b, \alpha) = \min_{\mathbf{w},b} f(\mathbf{w}, b) + \sum_{i=1}^{m} \alpha_i g_i(\mathbf{w}, b)$$

$$= \min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 + \sum_i \alpha_i \left(1 - y_i(\langle \mathbf{x_i}, \mathbf{w} \rangle + b)\right)$$

$$s.t. \quad \alpha \geq 0$$

$$\frac{\partial \mathcal{L}(\mathbf{w}, b, \alpha)}{\partial \mathbf{w}} = 0$$

$$\frac{\partial \mathcal{L}(\mathbf{w}, b, \alpha)}{\partial b} = 0$$

## Solving the dual problem. First solve:

$$\min_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b, \alpha) = \min_{\mathbf{w},b} f(\mathbf{w}, b) + \sum_{i=1}^{m} \alpha_i g_i(\mathbf{w}, b)$$

$$= \min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 + \sum_i \alpha_i \left(1 - y_i(\langle \mathbf{x_i}, \mathbf{w} \rangle + b)\right)$$

$$s.t. \quad \alpha \geq 0$$

$$\frac{\partial \mathcal{L}(\mathbf{w}, b, \alpha)}{\partial \mathbf{w}} = 0 \qquad \mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$$

$$\frac{\partial \mathcal{L}(\mathbf{w}, b, \alpha)}{\partial b} = 0 \qquad \sum_i \alpha_i y_i = 0$$

## Then replace:

$$\min_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b, \alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

## Then solve:

$$\max_{\alpha, \alpha_i \geq 0} \min_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b, \alpha) = \max_{\alpha, \alpha_i \geq 0} \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

$$s.t. \quad \alpha_i \geq a$$

$$\sum \alpha_i y_i = 0$$

After solving for $\alpha$ :

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$$

$$b = 1 - \langle \mathbf{x}_+, \mathbf{w} \rangle$$

## Then solve:

$$\max_{\alpha,\alpha_i \geq 0} \min_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b, \alpha) = \max_{\alpha,\alpha_i \geq 0} \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

$$s.t. \quad \alpha_i \geq a$$

$$\sum \alpha_i y_i = 0$$

After solving for $\alpha$ :

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$$

$$b = 1 - \langle \mathbf{x}_+, \mathbf{w} \rangle$$

For new points, predict:

$$\text{sign}\left(\langle \mathbf{w}, \mathbf{x}_j \rangle + b\right) = \text{sign}\left(\left\langle \sum_i \alpha_i y_i \mathbf{x}_i, \mathbf{x}_j \right\rangle + b\right) = \text{sign}\left(\sum_i \alpha_i y_i \langle \mathbf{x}_i, \mathbf{x}_j \rangle + b\right)$$

## Then solve:

$$\max_{\alpha,\alpha_i \geq 0} \min_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b, \alpha) = \max_{\alpha,\alpha_i \geq 0} \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

$$s.t. \quad \alpha_i \geq a$$

$$\sum \alpha_i y_i = 0$$

After solving for $\alpha$ :

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$$

$$b = 1 - \langle \mathbf{x}_+, \mathbf{w} \rangle$$

For new points, predict:

$$\text{sign}\left(\langle \mathbf{w}, \mathbf{x}_j \rangle + b\right) = \text{sign}\left(\left\langle \sum_i \alpha_i y_i \mathbf{x}_i, \mathbf{x}_j \right\rangle + b\right) = \text{sign}\left(\sum_i \alpha_i y_i \langle \mathbf{x}_i, \mathbf{x}_j \rangle + b\right)$$

## Kernels

A kernel is a function of two variables that can be written as dot product of the same feature map of these variables:

$$K(x, z) = \phi(x)^T \phi(z).$$

**Theorem (Mercer).** Let $K : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$ be given. Then for $K$ to be a valid (Mercer) kernel, it is necessary and sufficient that for any $\{x^{(1)}, \ldots, x^{(m)}\}$, $(m < \infty)$, the corresponding kernel matrix is symmetric positive semi-definite.

## Example kernel

$$\phi(x) = \begin{bmatrix} x \\ x^2 \\ x^3 \end{bmatrix}$$

$$k(x, y) = \langle \phi(x), \phi(y) \rangle = xy + x^2 y^2 + x^3 y^3$$

## Example kernel

$$K(x,z) = (x^T z)^2.$$

$$\begin{aligned} K(x,z) &= \left(\sum_{i=1}^{n} x_i z_i\right)\left(\sum_{j=1}^{n} x_j z_j\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j z_i z_j \\ &= \sum_{i,j=1}^{n} (x_i x_j)(z_i z_j) \end{aligned} \qquad \phi(x) = \begin{bmatrix} x_1 x_1 \\ x_1 x_2 \\ x_1 x_3 \\ x_2 x_1 \\ x_2 x_2 \\ x_2 x_3 \\ x_3 x_1 \\ x_3 x_2 \\ x_3 x_3 \end{bmatrix}.$$

# Survey

1

$$k_a(x,y) = \langle \phi_a(x), \phi_a(y) \rangle$$
$$k_b(x,y) = \langle \phi_b(x), \phi_b(y) \rangle$$

$$k(x,y) = k_a(x,y) + k_b(x,y)$$

is k a kernel?

yes $\quad \phi(x) = [\phi_a(x)^\top, \phi_b(x)^\top]^\top$

2

$$k_a(x,y) = \langle \phi_a(x), \phi_a(y) \rangle$$
$$k(x,y) = -k_a(x,y)$$

is k a kernel?

No, not PSD

3

$$k_a(x,y) = \langle \phi_a(x), \phi_a(y) \rangle$$
$$k(x,y) = c k_a(x,y) \quad c \geq 0$$

is k a kernel?

yes $\phi(x) = \sqrt{c}\phi_a(x)$

4

$$k(x,y) = x^\top A y$$
$$A \quad p.s.d.$$

is k a kernel?

Yes $A = LL^\top$, $\phi(x) = L^\top x$

5

Radial Basis Function (RBF) kernel

$$k(x,y) = \exp\left(-\frac{||x-y||^2}{\gamma}\right)$$

is k a kernel?

$$k(x,y) = \exp\left(-\frac{||x-y||^2}{\gamma}\right) = \exp\left(-\frac{||x||^2}{\gamma}\right)\exp\left(-\frac{||y||^2}{\gamma}\right) exp\left(\frac{2x^\top y}{\gamma}\right)$$
$$= \exp\left(-\frac{||x||^2}{\gamma}\right)\exp\left(-\frac{||y||^2}{\gamma}\right)\sum_{n=0}^{\infty}\frac{(x^\top y)^n}{n!}$$
$$= \exp\left(-\frac{||x||^2}{\gamma}\right)\exp\left(-\frac{||y||^2}{\gamma}\right)\langle [1, x, \frac{1}{\sqrt{2!}}x^2, ..., \frac{1}{\sqrt{i!}}x^i, ...], [1, y, \frac{1}{\sqrt{2!}}y^2, ..., \frac{1}{\sqrt{j!}}y^j$$

Yes

$\phi(x) = \exp\left(-\frac{||x||^2}{\gamma}\right)[1, x, \frac{1}{\sqrt{2!}}x^2, ..., \frac{1}{\sqrt{i!}}x^i, ...]$ Infinite dimensional feature map

## Learn non-linear boundaries

$$\text{sign}\left(\sum_i \alpha_i y_i \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\rangle + b\right) = \text{sign}\left(\sum_i \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}_j) + b\right)$$